EBIB Nr 5/2000 (14), Elektroniczne dokumenty. Artykuł

Marek Nahotko
Biblioteka Główna Politechniki Krakowskiej

Metadane

Problematyka dotycząca tworzenia metadanych to obecnie w świecie bibliotekarskim jedna z najbardziej rozwojowych dziedzin. Na ten temat powstaje sporo prac, jak na zagadnienia internetowe przystało - zwykle w formie cyfrowej, dostępnych na stronach WWW. Jak dotąd natomiast brak jest polskich publikacji poświęconych temu zagadnieniu. Artykuł ten jest próbą przybliżenia choć po części i bardzo ogólnie problemu metadanych polskim bibliotekarzom.

Metadane często nazywane są "danymi o danych" lub "informacją o informacji". Są to obrazowe określenia, ale mogą nieco zaciemnić sedno sprawy. W działalności informacyjnej termin "metadane" oznacza zdefiniowanie lub opis danych. Często przy tej okazji podaje się przykład katalogu bibliotecznego jako dobrze ustrukturyzowanego zbioru metadanych: każda karta katalogowa opisuje znacznie większy zasób informacji, którym jest skatalogowana książka lub inny dokument i odsyła użytkownika do tego dokumentu.

Podobnie metadane zawierają informacje o formie i treści dokumentów elektronicznych. Jedna z definicji mówi, że metadane to zwięzły i systematyczny zestaw informacji odsyłającej, który może być użyty do efektywnego i trafnego wyszukiwania większych zestawów informacji (czyli samych dokumentów elektronicznych). Metadane obejmują indeksowanie i katalogowanie wszelkich zasobów informacji w formie elektronicznej: stosowane są np. do opisu danych tekstowych, informacji o przestrzeni geograficznej, obrazów (dokumentów graficznych), muzyki (dokumentów dźwiękowych) i dzieł multimedialnych.

Metadane umożliwiają użytkownikom wyszukanie potrzebnej informacji w cyberprzestrzeni wraz z odpowiedzią na pytanie, w jakiej relacji pozostaje ona do innych informacji. Tradycyjne metody katalogowania nie są odpowiednie dla WWW: gwałtowny wzrost ilości danych uniemożliwia takie opracowanie. Natomiast wykorzystanie metadanych ułatwia zorganizowanie i zarządzanie informacją znajdującą się w WWW w sposób pozwalający na jej wyszukiwanie.

Bibliotekarze i pracownicy informacji naukowej rozumieją potrzebę standaryzacji w dziedzinie informacji, czego nie można powiedzieć o innych grupach użytkowników pracujących z metadanymi (np. autorzy, wydawcy). Ci pierwsi skupiają uwagę zarówno na tym, aby zapewnić odpowiednią formę informacji (jak zapisać informację - struktury danych), jak i jej treść (co zapisać). Ci drudzy koncentrują się jedynie na odpowiednim przygotowaniu struktur danych, co stanowi poważny mankament tych prac. Innymi słowy, skupiają się oni na tym, jakie pola należy utworzyć, bez zbytniej troski o to, jakie dane wpisać w te pola. Brakuje przy tym kartotek haseł wzorcowych i innych narzędzi kontroli terminologii, powszechnie i od dawna stosowanych w systemach bibliotecznych.

Jednym z najczęściej stosowanych formatów metadanych, którego inicjatorem w dużej mierze byli bibliotekarze, jest Dublin Core Metadata Element Set (DC). Utworzony on został dla dokumentów tekstowych w Web przez grupę związaną z OCLC **(http://purl.oclc.org/metadata/dublin_core/)**. Dublin Core to cały system różnych działań obejmujących warsztaty szkoleniowe, publikacje, grupy dyskusyjne i robocze pracujące na rzecz ułatwienia przeszukiwania zasobów elektronicznych. Efekt tych prac to piętnaście elementów Dublin Core (zob. rys. 1), które w intencji ich twórców powinny znaleźć szerokie zastosowanie i być łatwiejsze w użyciu niż biblioteczny katalog kartkowy.

Elementy Dublin Core mogą być podzielone na trzy klasy (Zawartość - Content, Własność Intelektualna - Intellectual Property, Dookreślenie - Instantiation) w następujący sposób:

Rys. 1. Schemat elementów Dublin Core

Zawartość (Content)	**Własność intelektualna (Intellectual Property)**	**Dookreślenie - Instantiation**
Tytuł (Title)	Twórca (Creator)	Data (Date)
Opis (Description)	Współtwórca (Contributor)	Format (Format)
Źródło (Source)	Własność (Rights)	Identyfikator (Identifier)
Język (Language)		
Relacja (Relation)		

Miejsce i Czas (Coverage)		

Część z tych elementów, takich jak tytuł, autor, opis rzeczowy jest dobrze znana wszystkim, którzy choć raz widzieli rekord bibliograficzny. Niektóre z elementów służą do zapisu danych technicznych i zawierają informację ważną dla dokumentów internetowych, jaką jest rozmiar pliku. Część jednak jest typowa dla nowych potrzeb cyberprzestrzeni i informuje np. kto posiada prawa do danego materiału.

- Title (Tytuł)
  Nazwa nadana dokumentowi przez Twórcę lub Wydawcę
- Creator (Twórca lub Autor)
  Osoba lub organizacja pierwotnie odpowiedzialna za stworzenie treści intelektualnych dokumentu. Np. są to autorzy w przypadku dokumentów drukowanych, artyści, fotograficy, ilustratorzy dla dokumentów audiowizualnych.
- Subject (Opis rzeczowy)
  Temat dokumentu. Zazwyczaj opis rzeczowy wyrażany jest za pomocą słów kluczowych lub wyrażeń określających przedmiot lub treść dokumentu. Planuje się użycie kontrolowanych słowników i schematów klasyfikacyjnych.
- Description (Opis)
  Tekst opisujący treść dokumentu, taki jak abstrakt w przypadku DLO (Document-like objects)[1] lub opis zawartości dla dokumentów wizualnych.
- Publisher (Wydawca)
  Organizacja odpowiedzialna za udostępnienie dokumentu w jego obecnej formie, taka jak wydawnictwo, instytucja sprawcza lub inne odmiany wydawców.
- Contributor (Współtwórca)
  Osoba lub organizacja nie zamieszczona w elemencie Twórca, która posiada istotny wkład intelektualny w powstanie dokumentu, lecz wkład ten jest wtórny w stosunku do osoby lub organizacji określonej w elemencie Twórca (np. redaktor, tłumacz lub ilustrator).
- Date (Data)
  Data udostępnienia dokumentu w obecnej formie. Rekomenduje się użycie 8-cyfrowej daty w formie RRRR-MM-DD. Możliwe jest użycie innej formy, jednak powinna ona być jednoznacznie zidentyfikowana.
- Type (Typ)
  Rodzaj dokumentu, taki jak strona domowa, powieść, poemat, dokument roboczy, raport techniczny, słownik. Dla zapewnienia przenoszalności Typ powinien być wybierany z listy, nad którą obecnie trwają prace.
- Format (Format)
  Format danych w dokumencie, wykorzystywany do identyfikacji oprogramowania oraz czasem sprzętu potrzebnego do wyświetlenia i działania na dokumencie. Podobnie jak Typ, Format także będzie wybierany z listy.
- Identifier (Identyfikator)
  Ciąg znaków lub numer używany do indywidualnej identyfikacji dokumentu. Przykładami dla zasobów sieciowych są URL i URN. Innymi powszechnie stosowanymi identyfikatorami są ISBN i ISSN.
- Source (Źródło)
  Ciąg znaków lub numer służący jednoznacznej identyfikacji dokumentu, z którego bieżący dokument pochodzi. Np. wersja PDF powieści w elemencie Źródło może zawierać ISBN powieści w formie książkowej, na podstawie której stworzono wersję PDF.
- Language (Język)
  Język lub języki, w których przedstawiona została intelektualna treść dokumentu. Dostępna jest pełna lista kodów języków.
- Relation (Relacja)
  Relacja pomiędzy dokumentem a innymi dokumentami. Element ten ma służyć wyrażaniu relacji istniejących pomiędzy dokumentami, które jednak istnieją samodzielnie. Np. obrazy (ilustracje) w dokumencie, rozdziały książki lub części pliku.
- Coverage (Miejsce i czas)
  Czasowe i/lub przestrzenne charakterystyki dokumentu.
- Rights (Własność)
  Opis praw autorskich, copyright, lub odesłanie do serwisu dostarczającego informacji o warunkach dostępności dokumentu.

DC posiada kilka pozytywnych cech, które dają nadzieję na jego dalszy szybki rozwój:

- Prostota - jest on prosty nawet dla nieprzygotowanego użytkownika;
- Spójność - dostarcza on spójne kategorie metadanych dla różnych typów dokumentów;
- Konsensus - DC ma charakter międzynarodowy, jest coraz powszechniej stosowany na wszystkich kontynentach;
- Elastyczność - może służyć do tworzenia zarówno prostych jak złożonych opisów;
- Dostosowawczość - wpisuje metadane w znane już i powszechnie zrozumiałe systemy, a więc może pracować w środowiskach już wcześniej powstałych i działających (takich jak tradycyjne biblioteki czy przeszukiwarki internetowe).

Twórcy Dublin Core blisko współpracują w zakresie standaryzacji metadanych z inną wpływową, aczkolwiek również nieformalną grupą, tzw. World Wide Web Consortium (w skrócie W3C - http://www.w3.org/). Nie znaczy to jednak, że poza tymi gremiami nie powstają inne schematy. Wręcz przeciwnie - każdy może stworzyć własny schemat, więc np. grupy związane ze sztuką wizualną lub modelowaniem obrazów wypracowały całkowicie odmienne zestawy elementów. Metadane są opracowywane dla bardzo różnych rodzajów dokumentów, jednak mogą one również być transponowane do innych schematów metadanych. Oznacza to, że np. pole <tytuł> w DC

może być przeniesione do pola tytułu w rekordzie formatu MARC. Te możliwości stanowią pewną nadzieję na ujednolicenie wielu schematów metadanych, opracowanych dla odmiennych potrzeb różnych dyscyplin.

W dokumencie WWW, w źródle pisanym w języku HTML elementy metadanych w standardzie DC wyglądają mniej więcej w następujący sposób (przykład dla tego artykułu):

Rys. 2. Przykład metadanych

```
<META NAME="DC.Title" CONTENT="Metadane">
<META NAME="DC.Subject" CONTENT="Metadane, Dublin Core, Przeszukiwarki">
<META NAME="DC.Description" CONTENT="Podstawowe informacje z zakresu metadanych">
<META NAME="DC.Publisher" CONTENT="Ossolineum">
<META NAME="DC.Creator" CONTENT="Marek Nahotko">
```

Olbrzymie i wciąż rosnące zasoby informacji i multimediów nazywane World Wide Web tworzone były bez specjalnej troski o zarządzanie informacją. Jednym z najważniejszych kroków w kierunku zmniejszenia chaosu panującego w nieuporządkowanym zbiorze informacji dostępnych w Web jest możliwość wprowadzenia porządku i standaryzacji w opisie ich treści, np. przy pomocy takich narzędzi jak DC, oraz odpowiedniego zastosowania tych narzędzi.

Informacja w WWW wyszukiwana jest przy użyciu tzw. przeszukiwarek, takich jak Yahoo, Lycos czy Alta Vista, poprzez wpisanie zapytania w formie słowa lub wyrażenia. Autor lub wydawca dokumentu internetowego umieszcza metadane (np. takie jak na rys. 2) na początku kodu HTML. Przeszukiwarka przegląda strony WWW na całym świecie porównując podane słowo lub wyrażenie z wpisanymi metadanymi. W efekcie otrzymuje się wykazy linków (adresów internetowych) dokumentów zarówno relewantnych, jak i (często) nierelewantnych, czasem w ilości tysięcy. Przeszukiwarki pracują w różny sposób, lecz zwykle ich wyniki wyszukiwania są bardzo mało trafne, a więc wyszukiwanie jest mało efektywne.

Użycie metadanych uporządkuje tę sytuację, szczególnie gdy możliwe będzie usunięcie synonimii i polisemii w terminach wyszukiwawczych przez wykorzystanie słowników kontrolowanych, znanych z innych zastosowań systemów komputerowych w bibliotekach. Ten pozytywny obraz komplikuje niestety fakt, że etykiety metadanych w HTML są często nadużywane dla uatrakcyjnienia stron komercyjnych, np. poprzez dodanie w tym miejscu terminów związanych z seksem. Dla uniknięcia takich błędów niektóre przeszukiwarki całkowicie ignorują etykiety <META> w HTML. Sytuacja może się zmienić, gdy metadane stosowane będą w sposób zapewniający ich poprawność i użyteczność. Elementy metadanych takie jak DC, oferują metodę organizacji i udostępniania informacji w WWW. W tym celu należy nie tylko utworzyć zestaw elementów, który będzie powszechnie akceptowany, ale także powszechnie stosowany.

Mówiąc o metadanych należy także wspomnieć o SGML i XML. Są to metajęzyki mogące zawierać elementy metadanych. SGML (Standard Generalized Markup Language) używa kodów ASCII do opisania tekstu. Każdy komputer rozumie ten kod. HTML, o którym wspominaliśmy już w kilku miejscach, jest bardzo prostym podzbiorem SGML. Przy pomocy SGML możliwe jest zidentyfikowanie i wyszukanie nazw, tytułów, opisów rzeczowych, połączeń pomiędzy tekstem i obrazami lub baz danych. Jednak SGML może być zrozumiany tylko przez specjalne przeglądarki, więc jeżeli ma być powszechnie dostępny, tłumaczony jest na HTML.

Nowym standardem w Web staje się obecnie XML (Extensible Markup Language), produkt W3C. O ile HTML to po prostu tekst ASCII zawierający etykiety umieszczane wg pewnych określonych reguł, to XML pozwala na określenie własnej struktury dokumentu. Standardem zapisu znaków w XML jest Unicode, co pozwala na unifikację wyglądu znaków narodowych, np. polskich, bez względu na użytą przeglądarkę, dzięki czemu dokument w języku polskim może być właściwie wyświetlony zarówno w kraju, jak i w Ameryce, co obecnie nie jest możliwe (lub trudne jest do osiągnięcia). XML jest narzędziem bardziej ogólnym od HTML. Ten ostatni zawiera informacje wyłącznie służące poprawnemu wyświetleniu danych na ekranie przy pomocy przeglądarki. Natomiast XML pozwala na logiczne uporządkowanie informacji wg potrzeb i żądań autora i/lub osób, z którymi wymieniamy dane.

Na podstawie tego, co napisano powyżej o metadanych i DC można by wyciągnąć wniosek, że nie jest to właściwie nic nowego, tym bardziej, że takie standardy jak MARC czy ISBD również pozwalają na opis dokumentów elektronicznych. Czy więc metadane to powielanie dawno już używanych narzędzi? Oczywiście, zdania na ten temat są podzielone. Już zestaw definicji przedstawiony na początku tego tekstu sugeruje, że metadane to co prawda katalogowanie, ale nie całkiem tradycyjne, a także nie całkiem katalogowanie. O sprawie tej ostatnio pisał m.in. Gradmann[2]. Zauważa on następujące różnice pomiędzy katalogowaniem a tworzeniem metadanych:

- Zbiory metadanych są inaczej używane niż katalogi biblioteczne (chodzi głównie o użycie internetowych przeszukiwarek i zwiększenie trafności wyszukiwania przy ich pomocy);
- Często nie są one tworzone przez zawodowych katalogerów;
- Dotyczą specjalnego rodzaju dokumentów (źródeł elektronicznych w WWW);
- Są tworzone z myślą o specjalnej grupie użytkowników (surfujących w cyberprzestrzeni);
- Istnieją poważne różnice pomiędzy stosunkiem metadanych i opisywanym źródłem elektronicznym a stosunkiem rekordu katalogowego i opisywaną książką.

Zatrzymajmy się dłużej przy ostatnim zagadnieniu. W skomputeryzowanej bibliotece istnienie rekordu opisu dokumentu w żaden sposób nie wpływa na użytkowanie samego dokumentu: rekord zawiera sam opis bibliograficzny oraz lokalne dane biblioteki (np. sygnatura). Dopiero znając sygnaturę można dotrzeć do dokumentu poprzez odrębny system udostępniania, często zresztą nie obywa się przy tym bez pomocy personelu biblioteki, który ręcznie dostarcza żądaną pozycję. Zwykle też niezbędna jest fizyczna obecność użytkownika w

bibliotece w momencie odbierania i zwrotu pozycji.

Zupełnie inaczej wygląda to w przypadku metadanych i dokumentów elektronicznych. Same metadane są częścią infrastruktury informacyjnej, pozwalając na bezpośredni dostęp do opisywanego dokumentu z dowolnego miejsca przyłączonego do sieci. Tak więc dotarcie do dowolnego elementu metadanych powoduje możliwość bezpośredniego i natychmiastowego dotarcia do treści samego dokumentu przez nie opisywanego, bez żadnych dodatkowych formalności ani niczyjego pośrednictwa. Oznacza to potrzebę tworzenia i utrzymania dodatkowych informacji, np. o oprogramowaniu używanym do odczytu danej aplikacji, czy adresie w sieci. To ostatnie jest zresztą przyczyną innego rodzaju kłopotów, dotyczących aktualności informacji o lokalizacji dokumentu elektronicznego w sieci; obecnie zagadnieniu różnego rodzaju identyfikatorów dokumentów elektronicznych, niezależnych od zmian w ich lokalizacji, poświęca się wiele wysiłków. W tym kontekście jest oczywiste, że błędny (np. nieaktualny) adres w sieci jest gorszy niż brak jakiegokolwiek adresu.

Interesujące są także możliwości znalezienia się w nowej sytuacji samych bibliotekarzy. Z pewnością muszą oni porzucić swoje przywiązanie do lokalnych katalogów opisujących lokalnie przechowywane dokumenty. Z drugiej jednak strony doświadczenia bibliotekarzy są nie do przecenienia również w zakresie metadanych. Mam tu na myśli głównie techniki i narzędzia zapewniające poprawność i jednolitość opisów, w tym różnego rodzaju kontrolowane słowniki, kartoteki haseł wzorcowych itp. Innym problemem jest zapewnienie jakościowej selekcji danych elektronicznych, tzn. mechanizmów pozwalających na ocenienie ich wartości merytorycznej. W obu przypadkach bibliotekarze mają wielowiekowe doświadczenia w rozwiązywaniu tych problemów. Z doświadczeń tych powinni korzystać także twórcy metadanych.

Przypisy:

[1] Termin *Document-like object* (DLO) powstał w 1995 r. Nie został on ściśle zdefiniowany - podaje się jedynie przykłady. Np. wersja elektroniczna artykułu z czasopisma lub słownika jest DLO, podczas gdy kolekcja przeźroczy nim nie jest. Intelektualną zawartością DLO jest głównie tekst.

[2] Gradmann Stefan: Cataloguing vs. Metadata: Old Wine in New Bottles? ICBC Vol. 28 Nr 4 s. 88-90.

Pierwotny adres: www.oss.wroc.pl/biuletyn/ebib14/nahotko.html
Adres w archiwum: ebib.oss.wroc.pl/arc/e014-02.html

Metadane / Marek Nahotko// W: Biuletyn EBIB [Dokument elektroniczny] / red. Bożena Bednarek-Michalska - Nr 5/2000 (14) czerwiec. - Czasopismo elektroniczne. - [Warszawa] : Stowarzyszenie Bibliotekarzy Polskich. KWE, 2001. - Tryb dostępu: http://www.ebib.pl/2001/14/nahotko.php. - Tyt. z pierwszego ekranu. - ISSN 1507-7187